

Limitations in Microelectronics – II. Bipolar Technology*

B. Hoeneisen and C.A. Mead

Abstract—The physical phenomena which will ultimately limit miniaturization of planar bipolar integrated circuits are examined. The maximum packing density is obtained by minimizing the supply voltage and the size of the devices. The minimum transistor size is determined by junction breakdown, punch-through, and doping fluctuations. For circuits that are fully active, the maximum number of circuit functions per chip is determined by power dissipation. The packing density of read-only memories becomes limited by the area occupied by devices and interconnections. The limitations of MOS and bipolar technologies are compared. It is concluded that read-only memories will reach approximately the same performance and packing density with MOS and bipolar technologies, while fully active circuits will reach the highest levels of integration with dynamic MOS or complementary MOS technologies.

1 Introduction

In this article, we examine the fundamental limitations of silicon planar bipolar integrated circuits. Limitations of planar MOS circuits were considered in Part I [1]. For both MOS and bipolar integrated circuits, the maximum number of circuit functions per unit area is determined either by power dissipation, or by the area occupied by transistors, interconnections and passive devices. For a given frequency of operation, reduction of the supply voltage and/or the circuit capacitances permits a reduction of power dissipation and interconnection area/transistor. Reducing the size of the devices not only reduces the area occupied by these devices, but also reduces the circuit capacitances. In addition, lower voltage devices can be made smaller. Thus, to maximize the circuit-packing density, it is necessary to minimize the supply voltage and the size of the devices.

The supply voltage cannot be lower than one diode drop ($\simeq 0.6V$). Otherwise, the transistors could not be turned on. The minimum supply voltage for proper circuit operation is typically 2 to 3 diode drops, depending on the circuit. For example, the minimum supply voltage of an RTL (resistor-transistor logic) gate is approximately two diode drops.

For a given supply voltage and doping concentration profile, the minimum transistor size is determined by punch-through, a condition where depletion regions overlap. To further reduce the device size, it is necessary to reduce the depletion thicknesses by increasing doping concentrations and reducing the supply voltage. The maximum doping concentrations are determined by junction field emission “breakdown.” (At high doping concentrations, the principal reverse-conduction mechanism of p - n junctions is tunneling of carriers across the junction. See [1].) Thus, a minimum-size transistor has its breakdown voltage equal to its punch-through voltage. Statistical fluctuations of the doping concentration can reduce the breakdown

*Re-typeset from original material by Donna Fox (August 2017). This work was supported in part by the Office of Naval Research and the General Electric Company. Originally published In *Solid-State Electronics*, Vol. 15, pp. 891–897, 1972.

or punch-through voltages. It is therefore necessary to set the supply voltage somewhat lower than the breakdown or punch-through voltage of the device.

It will be shown that the maximum packing density of fully active bipolar circuits is determined by power dissipation. The packing density of integrated circuits which are not fully active (e.g., read-only memories in which only a small fraction of the devices dissipate most of the power) becomes limited by the area occupied by transistors, interconnections, and passive devices. The analysis is necessarily approximate, since it requires a number of assumptions such as the geometry of the devices, the circuit configuration, and the maximum allowable power dissipation.

2 Breakdown and Punch-Through Limitation

An isoplanar bipolar transistor is shown in Fig. 1(a).

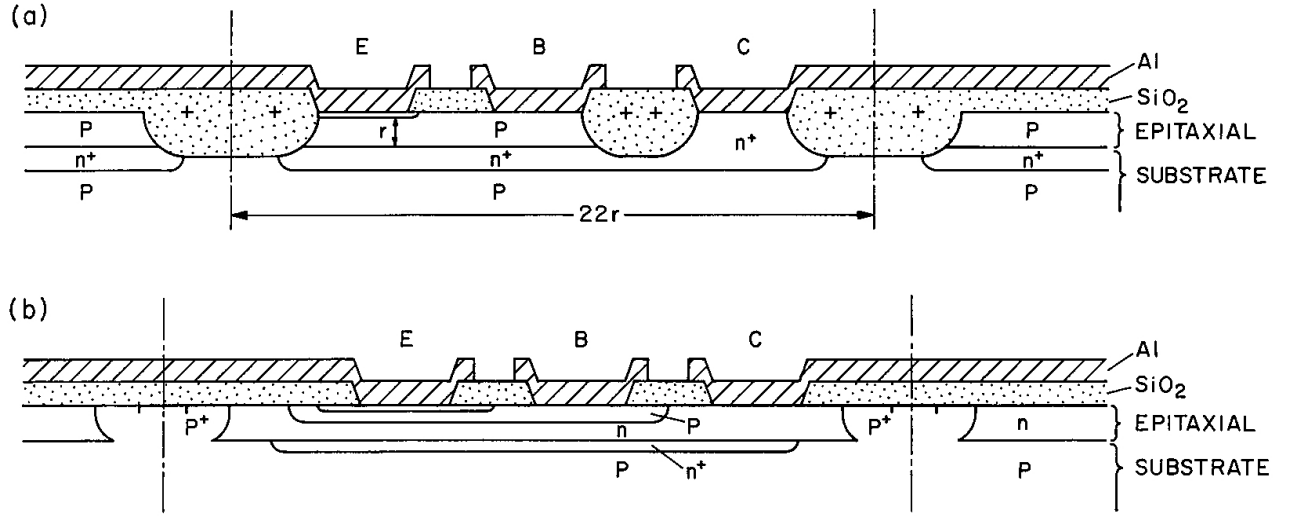


Figure 1: Approximate cross-section of minimum-size bipolar transistors. An isoplanar transistor is shown in (a), and a transistor with diffused isolation is shown in (b). The minimum-size isoplanar transistor occupies an area of approximately $22r \times 9r \approx 200r^2$, where r is the base thickness.

Since the base region is lightly doped compared to the emitter and collector, the depletion regions of the two junctions extend mainly into the base. For given voltages and base-doping concentration, the minimum-base region thickness is determined by punch-through, a condition where the depletion regions extending from the emitter and collector junctions overlap. The maximum base-doping concentration C_B is determined by collector junction “breakdown” as shown in Fig. 6 of Part I [1]. The minimum base thickness r will be set equal to $r_C + r_E$, where $r_C = \sqrt{(2\epsilon(V_{CB} + \phi)/qC_B)}$ is the collector junction-depletion thickness and $r_E = \sqrt{(2\epsilon\phi/qC_B)}$ is the emitter junction depletion thickness when the emitter is connected to the base. V_{CB} is the collector-base voltage, ϵ the silicon permittivity, ϕ the junction built-in voltage, and q the electronic charge. The minimum-base thickness r of the isoplanar transistor, determined by collector-junction breakdown and base punch-through, is presented in Fig. 2, curve A.

The maximum-substrate doping is equal to C_B , since both the substrate and base-doping concentrations are limited by collector-junction breakdown. If the substrate has a doping concentration C_B , the depletion region extending from the collector into the substrate has a thickness r_C . For a given base thickness, the minimum size of an isoplanar transistor can be determined approximately by geometrical considerations as

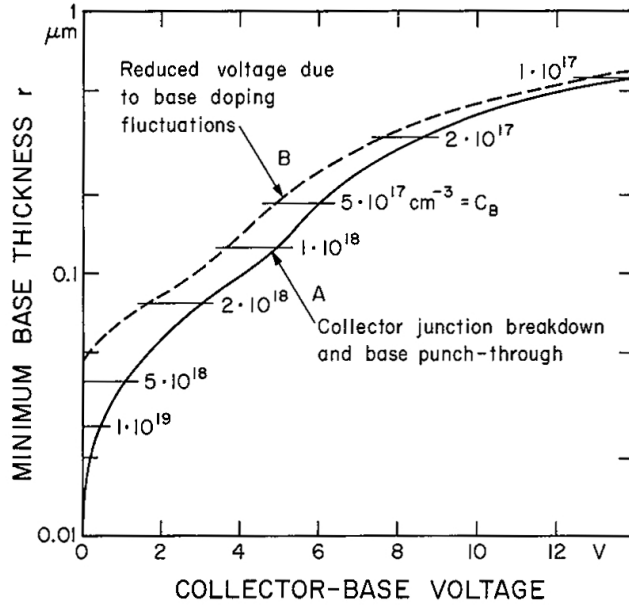


Figure 2: Solid line indicates the minimum-base thickness of an isoplanar transistor for a given collector-base voltage, determined by collector-junction breakdown and base punch-through. The collector-base voltage must be derated due to base-doping fluctuations, as indicated by the dashed line.

shown in Fig. 1(a). The isolation regions are broad enough to avoid punch-through between the collectors of adjacent transistors. The etch required to define the isolation regions is assumed isotropic, and the silicon dioxide layer is grown thick enough to avoid inversion of the underlying silicon. The minimum-size isoplanar transistor shown in Fig. 1(a), with a substrate-doping concentration equal to C_B , and with square emitter, base and collector contacts, occupies an area of $\simeq 200r^2$ and has a collector capacitance of $\simeq 350 \cdot \epsilon \cdot r$. If the circuit-packing density is limited by power dissipation rather than by the area occupied by the devices, it is convenient to reduce the collector capacitance by reducing the substrate doping concentration. The isolation regions must then be made broader to avoid punch-through between the collector regions of adjacent transistors. A transistor with the geometry shown in Fig. 1(a) except for broader isolation regions, with square emitter, base and collector contacts, and with a substrate-doping concentration equal to $\frac{1}{10}C_B$, occupies an area $\simeq 280r^2$, and has a collector capacitance of $\simeq 180 \cdot \epsilon \cdot r$. A planar bipolar transistor with diffused isolation is shown in Fig. 1(b). The area occupied by a minimum-size transistor of this type can be estimated by making several approximations. The emitter region is degenerate, and has a doping concentration of $\simeq 10^{21}/\text{cm}^3$. The doping concentration of the base, which is made lower than that of the emitter to insure good emitter efficiency, can be as high as $\simeq 10^{19}/\text{cm}^3$. We make the simplifying assumptions that the doping concentration of the epitaxial collector region is considerably lower than that of the base, and that the collector-junction depletion region does not reach the n^+ buried collector. For a given collector-base voltage, the maximum collector-doping concentration C_C determined by collector-junction breakdown can then be obtained directly from Fig. 6 of Part I[1]. The collector-junction depletion thickness is $r_C \simeq \sqrt{(2\epsilon(V_{CB} + \phi)/qC_C)}$. With this depletion thickness, the minimum size of a transistor with diffused isolation can be estimated using geometrical considerations as shown in Fig. 1(b). The result is that a minimum-size transistor with diffused isolation occupies $\simeq 2.1$ times more area, and has a collector capacitance $\simeq 2.5$ times greater than a minimum-size isoplanar transistor with the same voltage rating, contact area, and substrate-doping concentration. The increased area and capacitance of the transistor with diffused isolation is mainly due to the separation between the base and isolation diffusion which is required to avoid punch-through between these regions. In what follows, only the isoplanar transistor is considered.

3 Doping Fluctuation Limitation

As devices are made smaller, the number of dopant atoms in a characteristic volume of the device decreases until its statistical fluctuations become important. Consider a minimum-size isoplanar transistor designed to have a collector-junction breakdown voltage equal to the base punch-through voltage V_{CB} , as in the previous section. The effect of base-doping fluctuation is to decrease the breakdown voltage (if C_B increases), or decrease the punch-through voltage (if C_B decreases). Thus, due to base-doping fluctuations, the maximum allowable collector-base voltage must be derated. A chip with 10^6 isoplanar transistors will be considered. We shall arbitrarily choose the maximum allowable collector-base voltage V_{CBM} in such a way that, with an 80-percent certainty, none of the 10^6 transistors will have a collector breakdown or punch-through voltage lower than V_{CBM} .

The amount by which the collector-base voltage V_{CB} must be derated due to base-doping fluctuations can be estimated as follows: The breakdown, or punch-through voltages are altered significantly only if the base-doping fluctuation causes a doping concentration variation in a volume $\simeq r_C^3$ or greater; r_C is the collector-junction depletion thickness. The collector-junction depletion region of a minimum-size isoplanar transistor, such as the one shown in Fig. 1(a), can typically be divided into 30 cubes of volume r_C^3 . The doping fluctuation ΔC_B which, with an 80-percent certainty, is not exceeded in any of the 30×10^6 cubes of volume r_C^3 is given by

$$\Delta C_B \cong 5.8 \cdot \frac{\sqrt{(n)}}{r_C^3} = 5.8 \cdot \sqrt{\left(\frac{C_B}{r_C^3}\right)}. \quad (1)$$

This calculation assumes that the number of dopant atoms in a volume r_C^3 has a Gaussian distribution with mean $n \equiv C_B r_C^3$, and standard deviation $\sqrt{(n)}$. The result does not depend strongly on the number of cubes assumed. The maximum allowable collector-base voltage V_{CBM} is determined either by breakdown (with a base-doping concentration $C_B + \Delta C_B$), or by punch-through (with a doping concentration $C_B - \Delta C_B$), whichever is smaller, and is presented in Fig. 2, curve B. Notice that the maximum collector-base voltage determined by breakdown and punch-through must be derated by about 1.4V due to base-doping fluctuations. In a practical design, the collector-base voltage must be derated further due to the manufacturing tolerances of the base-doping concentration.

4 Power Dissipation Limitation

All bipolar transistor circuits have some sort of current-limiting devices, such as resistors and current sources. These current-limiting devices can usually not be avoided, even in complementary circuits, due to the low impedance of the base-emitter junctions. The power dissipation of a circuit can be divided into static power dissipation (associated with the steady-state currents) and dynamic power dissipation (associated with transient current which charge and discharge the circuit capacitances). The dynamic power dissipation is proportional to frequency. The static-power dissipation of digital circuits is determined primarily by the current-limiting devices in the circuit, and not by the active transistors, since these are either on or off. The resistors can be chosen to obtain the desired static power-dissipation density. These resistors and the circuit capacitances then determine the circuit-time constants, which in turn set an upper limit to the frequency of operation. The maximum frequency of operation of bipolar integrated circuits is therefore determined either by dynamic power dissipation, by the circuit-time constants which depend on the static-power dissipation, or, in saturating circuits, by the lifetime of minority carriers in the base.

Let us examine the simple circuit shown in Fig. 3.

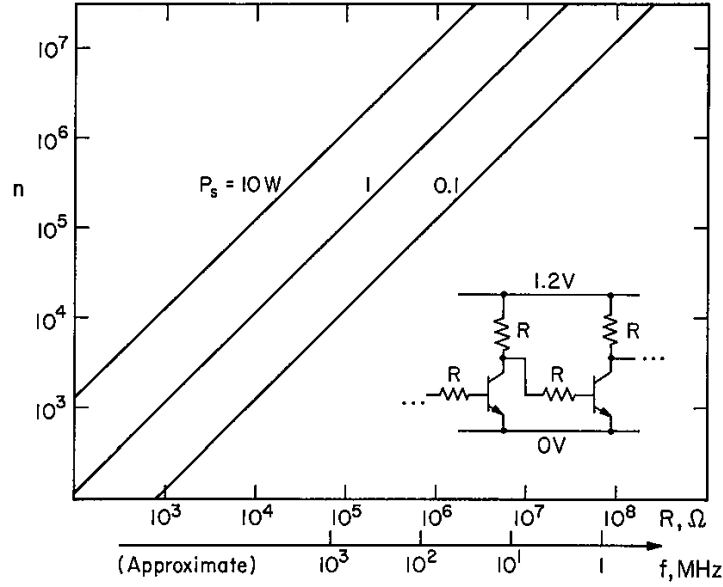


Figure 3: For a given frequency of operation f and static-power dissipation P_s , the maximum number of inverters n and resistor value R are determined for the particular circuit indicated. At the frequency f , the dynamic-power dissipation is $P_d \simeq 0.36 P_s$. It is assumed that the equivalent capacitance of the resistors is equal to the collector-junction capacitance of the transistors (i.e., $180 \cdot \epsilon \cdot r$ with $r = 0.07 \mu\text{m}$).

The circuit consists of a series of resistor-transistor logic (RTL) inverters connected in cascade (fan out = 1), so that half of the inverters are in the high-level state, and half are in the low-level state. The supply voltage is 1.2V, so the minimum base thickness of the isoplanar transistors is $r = 0.07 \mu\text{m}$ as shown in Fig. 2. If the substrate doping concentration is equal to $\frac{1}{10}$ th of the base-doping concentration, the minimum-size isoplanar transistor occupies an area of $\simeq 280r^2$, and has a collector capacitance of $\simeq 180 \cdot \epsilon \cdot r$, as indicated in a previous section. We assume that each resistor (including interconnections) occupies the same area as a transistor, so that each inverter occupies $\simeq 840r^2$. This area corresponds to a maximum density of $\simeq 2.4 \times 10^7$ inverters/cm². At this packing density, the value R of the resistors required to have a static power dissipation density of 1 W/cm² is $\simeq 20 \text{ M}\Omega$. The rise time of the collector of a transistor that is turned off is approximately $\tau = RC$, where C is the total capacitance associated with the collector node, i.e., collector-junction capacitance plus capacitance of the two resistors connected to the collector. In our example $\tau \simeq 77 \text{ nsec}$, if we assume that the equivalent capacitance of each resistor is equal to the collector junction capacitance. Assuming that the recovery time of the transistor is shorter than τ , the maximum frequency of operation is determined by the collector rise time, and is of the order of $\frac{1}{5}\tau \simeq 2.6 \text{ MHz}$. The dynamic power dissipation density at 2.6 MHz is approximately 0.36 W/cm².

This example illustrates two important limitations of densely-packed bipolar integrated circuits. First, the resistor values required to have a reasonable static power-dissipation density are exceedingly large. Second, the maximum frequency of operation is limited to quite low values by both the circuit-time constants and by dynamic power dissipation. Notice that at the maximum frequency $f = \frac{1}{5}RC$, the dynamic-power dissipation of n inverters is $P_d = nfCV^2 = nV^2/5R$, which is approximately 0.36 P_s for the circuit considered above, and is independent of C . P_s is the static-power dissipation of n inverters.

For a given frequency of operation f and maximum-power dissipation/chip P , the maximum number of inverters/chip n , and the resistor value R , can be determined from Fig. 3 for the particular circuit considered.

We conclude that the minimum-power dissipation/circuit function is determined by the required frequency of operation, and by the supply voltage and circuit capacitances. This power dissipation limits the number of circuit functions/chip in fully active bipolar circuits. In circuits which are not fully active, such as read-only memories in which only a small fraction of the devices dissipate most of the power, the area occupied by the devices and interconnections becomes the limiting factor.

5 Metal Migration Limitation

Metal migration is an important reliability consideration in integrated circuit design. To avoid “strip burn out,” the instantaneous current density in aluminum conductors should be kept substantially below 10^6 A/cm^2 [2]. Ohmic drop must also be considered. At 10^6 A/cm^2 , the ohmic drop in aluminum conductors is $\simeq 3 \text{ V/cm}$. Metal migration does not limit transistor size, but rather limits the number of circuit functions per unit area. We shall see that metal migration and power dissipation are two closely-related limitations.

Let us consider the supply lines. For a given power dissipation P and supply voltage V_{CC} , the total average current that must be carried by the supply lines is $I = P/V_{CC}$. Therefore, for a given power dissipation, supply voltage and conductor-current density, the total supply line cross-section area is independent of circuit complexity. Maintaining the present-day power-dissipation density and metallization thickness ($\simeq 1 \mu\text{m}$), we can expect the relative chip area occupied by supply lines to remain approximately independent of circuit complexity. For a given clock-pulse rise and fall time constant τ , the maximum instantaneous clock-line current is $I \simeq CV_{CC}/\tau$, where C is the load capacitance. If τ is, say, $1/10^{\text{th}}$ of a cycle, the total instantaneous current in the clock lines is $I \simeq 10 P_d/V_{CC}$, where P_d is the dynamic-power dissipation associated with C . The total current is again proportional to power dissipation, so we can expect that the relative chip area occupied by clock lines will also remain approximately independent of circuit complexity, provided that the metallization thickness is not reduced.

As a particular example, we shall consider a 1 cm^2 chip with 10^6 minimum-size inverters, a supply voltage of 1.2 V , a static-power dissipation of 1 W , and a dynamic-power dissipation of 0.36 W at 100 MHz , as shown in Fig. 3. We assume that a supply line is connected to 2×10^3 inverters. The maximum mean current in this line is 2.3 mA . If a maximum current density of 10^5 A/cm^2 is allowed, the minimum conductor cross-section area required would be $2.3 (\mu\text{m})^2$. For a metallization thickness of $1 \mu\text{m}$, all supply lines would occupy an area of approximately 0.23 cm^2 .

6 High-Valued Resistors

To improve the yield, it is convenient to use the same processing steps to make the transistors and resistors in the integrated circuit. The highest sheet resistance is obtained by using the base region of the isoplanar transistors. Let us consider a minimum-size 1.2 V transistor with a base-region thickness of $r = 0.07 \mu\text{m}$, and a base-doping concentration of $2.3 \times 10^{18}/\text{cm}^3$, as shown in Fig. 2. The $1 - V$ characteristic of a one-square pull-up resistor is calculated using field-effect transistor equations, and is presented in Fig. 4. The current at $v = 0\text{V}$ is $\simeq 5 \times 10^{-5} \text{ A}$, which corresponds to a linear resistor of $\simeq 24 \text{ k}\Omega/\text{square}$.

Alternative high ohm/square resistors are MOS resistors and cermet (ceramic-metal) thin-film resistors. Sheet resistances of $7 - 25 \text{ k}\Omega/\text{square}$ are obtained with good controllability by using non-saturated MOS transistors as resistors [3]. For example, the minimum-size MOS pull-up transistor considered in Fig. 4

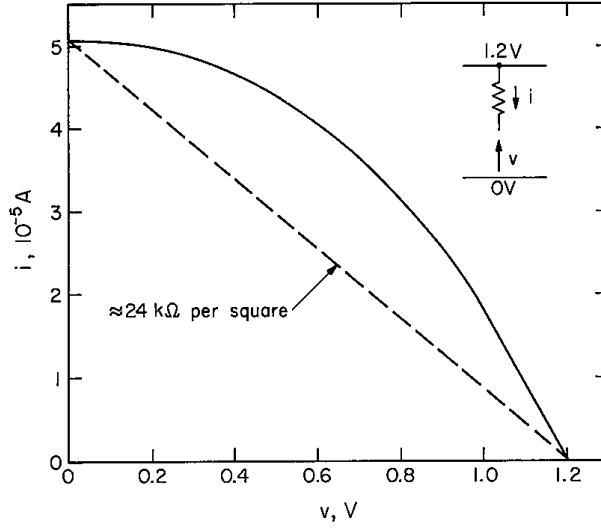


Figure 4: $i - V$ characteristic of the base region of a minimum-size 1.2 V isoplanar transistor used as a pull-up resistor. The emitter and collector regions are connected to 1.2 V. The base width and length are equal.

of Part I [1] has a sheet resistance of $\simeq 9 \text{ k}\Omega/\text{square}$. Cermet resistors are manufactured with sheet resistances up to $\simeq 10 \text{ k}\Omega/\text{square}$.¹

The maximum attainable sheet resistance is an important limitation of fully active bipolar circuits. Let us again consider the circuit shown in Fig. 3. The maximum packing density is obtained by increasing the length-to-width ratio of the resistors until the densely-packed chip has an acceptable power dissipation density. The base region of a minimum-size 1.2 V isoplanar transistor has a sheet resistance of $\simeq 24 \text{ k}\Omega/\text{square}$ as shown in Fig. 4. The minimum-width resistor occupies $\simeq 6(\mu\text{m})^2$ of chip area per $\text{M}\Omega$. The static-power dissipation density of a densely-packed chip is $1 \text{ W}/\text{cm}^2$ if the length-to-width ratio of the resistors is chosen to be $\simeq 110$. In this case, most of the chip area is occupied by the resistors. The packing density is then $\simeq 6.5 \times 10^6 \text{ resistors}/\text{cm}^2$, and the maximum frequency of operation is $\simeq 2 \text{ MHz}$.

7 Conclusions

For a given collector voltage, the minimum base thickness of an isoplanar transistor is determined by collector-junction breakdown, base punch-through, and base-doping fluctuations, as shown in Fig. 2. A minimum-size 1.2 V transistor has a base thickness of $0.07 \mu\text{m}$, and a base-doping concentration of $2.3 \times 10^{18}/\text{cm}^3$. For a given base-region thickness r , the minimum area of an isoplanar transistor, determined by geometry considerations, is approximately $200 r^2$.

The required frequency of operation and the circuit capacitances determine an upper limit to the resistor values. These resistors and the supply voltage determine the minimum-power dissipation per circuit function. For a given frequency of operation, the power dissipation per circuit function is therefore minimized by minimizing the supply voltage and the circuit capacitances. The maximum allowable power dissipation/chip then sets an upper limit to the number of circuit functions/chip. An upper limit to the packing density is determined by the area occupied by transistors, resistors, and interconnections. The minimum area occupied by the interconnections is determined by metal migration.

¹Private communication.

If the resistors and transistors are to be made with the same manufacturing steps, it is convenient to use the base region of the isoplanar transistors as depletion-mode resistors. Then, sheet resistances of $10 - 30 \text{ k}\Omega/\text{square}$ are obtained. As an example, let us consider a fully active circuit with a supply voltage of 1.2 V as shown in Fig. 3. If base resistors with a length-to-width ratio of 10 are used, a static power-dissipation density of $\simeq 1 \text{ W}/\text{cm}^2$ is obtained with $\simeq 3 \times 10^5$ inverters/ cm^2 (i.e., 9×10^5 devices/ cm^2). Taking into account the parasitic capacitances, the maximum frequency of operation of this circuit is $\simeq 150 \text{ MHz}$. At 150 MHz , the total power dissipation is $\simeq 1.4 \text{ W}/\text{cm}^2$.

The power dissipation per transistor in read-only memories is low because there are many driver transistors/pull-up resistors. The maximum-packing density of read-only memories is therefore limited by the area occupied by the devices and interconnections. Furthermore, since the transistors have common collector and base regions, extreme packing densities can be achieved. For example, a 1.2 V read-only memory with isoplanar transistors can have up to 1×10^8 transistors/ cm^2 .

8 Comparison of MOS and Bipolar Technologies

The area occupied by a minimum-size 1.2 V MOS driver transistor with a channel length-to-width ratio of 1 is $\simeq 0.6 (\mu\text{m})^2$, and its gate plus drain junction capacitance is $\simeq 4 \times 10^{-4} \text{ pF}$ [1]. The area occupied by a minimum-size 1.2 V isoplanar bipolar transistor with a substrate-doping concentration equal to $1/10^{\text{th}}$ the base-doping concentration is $\simeq 1.4 (\mu\text{m})^2$, and its collector capacitance is $\simeq 1.3 \times 10^{-3} \text{ pF}$. For minimum-size devices of equal voltage rating, the MOS transistor has a factor of 2–3 advantage in area and capacitance over the bipolar transistor.

The maximum-packing density, and the corresponding maximum frequency of operation of several bipolar and MOS transistor circuits are presented in Fig. 5 (Table 1).

	Isoplanar bipolar		Static non-complementary MOS		Dynamic MOS	
	Density	Frequency	Density	Frequency	Density	Frequency
1.2 V inverters with 240 k Ω diffused or MOS resistors	$9 \times 10^5^*$	150†	$9 \times 10^5^*$	200†	—	—
2 V fully dynamic inverters	—	—	—	—	$3 \times 10^7\ddagger$	30§
1.2 V read-only memory¶	$1 \times 10^8\ddagger$	0.5**	$1 \times 10^8\ddagger$	0.5**	—	—

*Determined by static power dissipation $P_s = 1 \text{ W}/\text{cm}^2$.

†Determined by the circuit time constants.

‡Determined by the area occupied by transistors and interconnections.

§Determined by dynamic power dissipation $P_d = 2.1 \text{ W}/\text{cm}^2$ and/or metal migration.

¶Two level metallization assumed. 1 cm^2 circuit.

**Determined by metal migration and/or the circuit time constants.

Figure 5: Table 1. Maximum-packing density (devices/ cm^2) and corresponding maximum frequency of operation (MHz) of several bipolar and MOS transistor-integrated circuits.

The packing density of fully active bipolar and static MOS circuits is limited by power dissipation. For these circuits, the maximum-packing density and frequency of operation are approximately the same for

both bipolar and MOS technologies. Read-only memories have many driver transistors connected to one pull-up resistor. The packing density of read-only memories is therefore not limited by power dissipation, but by the area occupied by devices and interconnections. Furthermore, due to the simplicity of the circuits, extreme packing densities can be achieved as shown in Fig. 5.

We conclude that read-only memories will reach approximately the same performance and packing density with MOS and bipolar technologies, while fully-active circuits will reach the highest levels of integration with dynamic MOS or complementary MOS technologies.

References

- [1] B. Hoeneisen and C.A. Mead. *Solid-State Electronics*, Vol. 15 (7), pp. 819–829, (1972).
- [2] I.A. Blech and E.S. Meieran. *Appl. Phys. Lett.*, Vol. 11, p. 263, (1967).
- [3] L. Vadasz. *IEEE Trans.*, ED-13, p. 459, (1966).